Pattern Recognition Letters 49 (2014) 201–206

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Real-time local stereo via edge-aware disparity propagation ☆

Xun Sun [a,b,*], Xing Mei [c], Shaohui Jiao [b], Mingcai Zhou [b], Zhihua Liu [b], Haitao Wang [b]

[a] Baidu Institute of Deep Learning, Beijing, China
[b] China Lab, Samsung Advanced Institute of Technology, Beijing, China
[c] NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing, China

ARTICLE INFO

Article history:
Received 24 December 2013
Available online 1 August 2014

Keywords:
Stereo matching
Disparity propagation
Edge-preserving smoothing

ABSTRACT

This letter presents a novel method for real-time local stereo matching. Different from previous methods which have spent many efforts on cost aggregation, the proposed method re-solves the stereo problem by propagating disparities in the cost domain. It is started by pre-detecting the disparity priors, on which a new cost volume is built for disparity assignment. Then the reliable disparities are propagated via filtering on this cost volume. Specially, a new $O(1)$ geodesic filter is proposed and demonstrated effective for the task of edge-aware disparity propagation. As can be expected, the proposed framework is highly efficient, due to leaving double aggregation on left–right views, as well as costly post-processing steps, out of account. Moreover, by properly designing a quadric cost function, our method could be extended to good sub-pixel accuracy with a simple quadratic polynomial interpolation. Quantitative evaluation shows that it outperforms all the other local methods both in terms of accuracy and speed on Middlebury benchmark. It ranks 8th out of over 150 submissions if sub-pixel precision is considered, and the average runtime is only 9 ms on a NVIDIA GeForce GTX 580 GPU.

1. Introduction

Dense two-frame stereo is one of the most extensively studied topics in computer vision [14]. It takes in a rectified image pair captured from two viewpoints, and aims at inferring the depth information in the form of disparity. Generally, different stereo algorithms could be broadly separated into two major classes: *local* and *global* methods.

Local methods compute each pixel's disparity independently. They aggregate the cost (support) from a local region (usually a window) with an implicit smoothness assumption, and choose the disparity hypotheses with the minimal cost (Winner-Take-All). On the contrary, global methods optimize the disparities of all pixels at the same time. They formulate the stereo problem with an explicit smoothness assumption, and infer the disparity map by minimizing a global energy function. To approximate the energy minimum, powerful global optimizers such as belief propagation [3,23] and graph cuts [2] are used. The adoption of these time-consuming optimizers make these global methods far away from being real-time.

Global methods were generally expected to be more accurate than local methods until the use of modern edge-preserving filters [16,11,4]. Yoon and Kweon [25] firstly demonstrated that, by aggregating costs with a joint bilateral filter, the local methods could produce good results on par with those generated by global methods. Although a full-kerneled implementation of bilateral filtering is slow, since that time adaptive support-weight approach has become popular in the stereo community, and various great improvements on weighting scheme are continuously being made [13,12,19,20,8]. Current leading local methods can produce high-quality disparity map with sharp edges, and usually are with a kernel independent of any window size thus very efficient [6]. All of the above works have greatly advanced the state-of-arts. However, their efforts mainly focus on the cost aggregation step, and the computational redundancy still remains. Recalling the recent state-of-art local methods, nearly all of them share such a same work flow: firstly the matching costs on the left and right views are aggregated separately. Then with the WTA results on both views, a left–right consistency check is employed to remove the occluded pixels. Finally these "missing" pixels are filled with a background interpolation followed by a weighted-median filter. Even with acceleration, the adaptively-weighted median filter (e.g., with bilateral weights [13]) is time-consuming (with a complexity comparable to the cost aggregation). As a result, to calculate a final disparity map on the reference view (left view), this

☆ This paper has been recommended for acceptance by L. Yin.

* Corresponding author at: Baidu Institute of Deep Learning, Beijing, China. Tel.: +86 10 18612940083.
E-mail address: xun.s@qq.com (X. Sun).

http://dx.doi.org/10.1016/j.patrec.2014.07.010
0167-8655/

popular pipe-line would take *three* expensive operations in aggregation-level.

Recently, Ma et al. proposed to directly apply a weighted median filter on a noisy disparity map [7]. The authors claimed that, if the inliers are dominant in local regions, only one aggregation-level operation by weighted median filtering could suffice to remove outliers while preserving edges. Nevertheless, there are non-trivial errors that could not be recovered in such a mechanism: to generate the initial disparity map, they use simple box-aggregation and background interpolation to fill in the holes after left–right check. This is questionable since not only occluded pixels fail in this test, a large portion of other pixels, e.g., pixels in textureless regions could fail as well. The main reason of this problem is the poor matching quality of box-aggregation (comparing to the sophisticated methods). Therefore, their method could lead to "blur" results when mis-matched pixels dominate the local regions. Even median filter cannot eliminate this kind of errors.

In this letter, the local stereo problem is re-examined, and a novel framework with only one aggregation-level operation is proposed. At first, inspired from recent insights [9], with the box-filtered initial matching costs, a set of stable pixels are determined by a left–right consistency check. Besides, for each stable pixel, a compact sub-set for its disparity varying range is detected from the cost profile. A new cost volume is then built with the above two key ingredients. More concretely, the cost for a pixel is re-computed according to its membership (stable or unstable). For a stable pixel, a penalty is put on disparity variation departing from its pre-detected disparity priors. In contrast, for an unstable pixel, the cost will keep zero for all disparity hypotheses. Thus an edge-aware filter applied on this new cost volume leads to reliable disparity propagation in cost domain: for the unstable pixels, the cost for their disparity selection will purely depend on the stable pixels. Comparing to [7], our solution delivers sharp disparity maps in a unified framework with no early hard decision. Both the occluded and mis-matched pixels are considered as unstable pixels, and their disparities are assigned by softly aggregating support from stable pixels. Comparing to [17], a previous global method which models the regularization from ground control points into a energy term, our method achieves real-time performance and do not need a sophisticated mechanism for GCPs' generation. Another great advantage of our method over discrete method such as [7] is that, with a specially designed quadratic cost term, our method could be extended to good sub-pixel accuracy using a simple quadratic polynomial interpolation.

To avoid prorogating disparities across depth edges, a new $O(1)$ geodesic filter inspired from [19] is proposed in this letter. Comparing to the original MST (minimal spanning tree) based filter, this new filter has two important advantages: firstly it is more GPU-friendly since it treats each scanline as a separate tree. Secondly, it makes use of every edge on the 4-connected image grid hence no information is lost as in a MST structure. These merits combined make our filtering scheme not only more efficient, but also more accurate than [19]. There are other image grid based filtering techniques such as [12,20,18], however, recursive filter based methods [12,20] suffer from a directional bias (see Section 2.1 [18]) needs an additional high-dimensional buffer for saving the temporal results in the case of cost-volume filtering (in contrast, our filtered values are updated in place).

Experimental results demonstrate the effectiveness of our approach. It is current top performer both in terms of accuracy and speed on Middlebury benchmark. In summary, the contributions in this work are:

- A novel and simplified framework for local stereo matching with leading accuracy and real-time performance.
- A $O(1)$ geodesic filter proposed for disparity propagation. This filter is very effective for stereo matching.

2. Algorithm

This section presents the proposed stereo matching algorithm. In Section 2.1, the stereo framework by propagating reliable disparities is presented. Then in Section 2.2, a new $O(1)$ geodesic filter is presented for the purpose of preserving structures in disparity propagation (Section 2.1).

2.1. Local stereo via reliable disparity propagation

The framework of the proposed reliable disparity propagation scheme is shown in Fig. 1. First, a set of stable pixels, as well as their disparity sub-sets (candidates with high likelihood to be correct) are determined in a pre-processing step. Second, a new cost volume is built purely with the detected disparity priors. Then an edge-preserving filter is applied on this cost volume for disparity propagation. Finally, the disparity map is calculated by WTA optimization. In the following, the detailed descriptions of the pre-processing step and disparity propagation procedure are presented.

Pre-processing. In this step, all the image pixels are roughly divided into stable or unstable pixels. Then for each stable pixel, a compact disparity sub-set is extracted.

Given a pixel $\mathbf{p} = (x, y)$ in I^{Left}, a matching cost $C(\mathbf{p}, d)$ at disparity d is initialized with the exact measure used in [13,19,12]. To suppress the noises, a 5×5 box filter is applied on the initial cost volume. This is done for all disparity levels, but its complexity is trivial since such a small-kerneled filter is very efficient in GPU implementation (the average runtime is only 1 ms on Middlebury data sets, which is negligible comparing to other operations). With the initial cost, a raw disparity map D^{Left}_{Raw} on left view is calculated by WTA. Symmetrically, a corresponding disparity map D^{Right}_{Raw} is also computed. Then a left–right consistency check is employed to roughly divide all image pixels into stable or unstable pixels. For a stable pixel $\mathbf{p}$, its disparity value d on D^{Left}_{Raw} should strictly equal to $D^{Right}_{Raw}(\mathbf{p} - (d, 0))$.

Recently, Min et al. [9] proposed a compact representation for local stereo matching. In their approach, a per-pixel sub-set of disparity searching range is extracted by box-filtering and cost profile analysis. In this letter, a simplified compact representation is used and it is proven to be still very effective. Instead of sorting the local

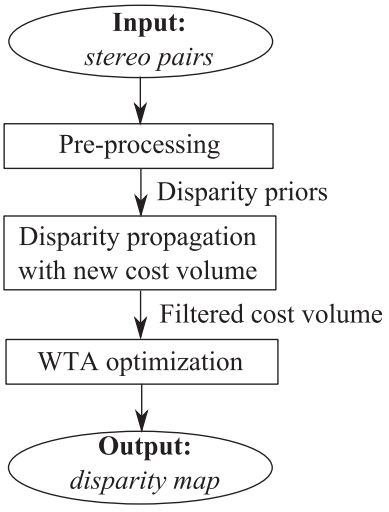

Fig. 1. Block diagram of our stereo matching algorithm.